(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2000－148788
（P2000－148788A）
(43)公開日　平成12年５月30日(2000.5.30)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/40	３７０Ｂ	５Ｂ００９
17/27		Ｇ０６Ｋ　9/20	３４０	５Ｂ０２９
Ｇ０６Ｋ　9/20	３４０	Ｇ０６Ｆ　15/20	５５０Ｆ	５Ｂ０７５
		15/401	３１０Ａ	

審査請求　未請求　請求項の数11　ＦＤ　（全 9 頁）

(21)出願番号　特願平10－328806

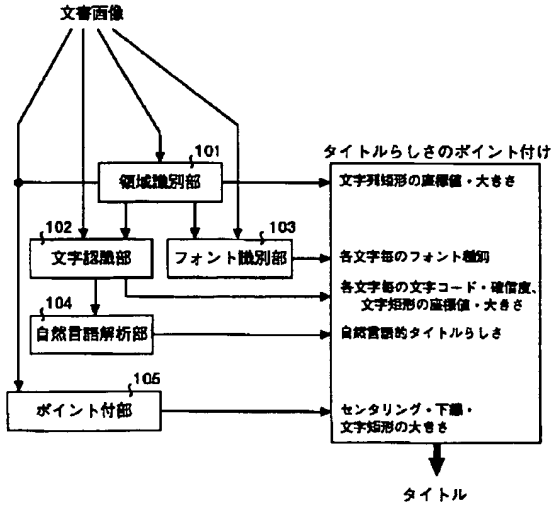
(22)出願日　平成10年11月５日(1998.11.5)

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　大内　茂樹
東京都大田区中馬込１丁目３番６号　株式会社リコー内

Ｆターム(参考)　5B009 QA12
5B029 AA01 BB02 CC27
5B075 ND03 NK02 NK04 NK32 NK39
UU06

(54)【発明の名称】　文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法，並びに文書検索方法

(57)【要約】
【課題】　特定の文書形式に依存せずにタイトル固有の特徴をポイントとして用いることにより，ポイント数の多い文字列領域をタイトルとして自動抽出し，タイトル抽出の的確性および文書検索時の利便性を向上させること。
【解決手段】　領域識別部１０１で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識部１０２と，上記文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別部１０３と，文字認識部１０２の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析部１０４と，上記文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付部１０５と，を備えた。



【特許請求の範囲】

【請求項１】　画像入力装置から入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出装置において，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識手段と，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と，前記文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と，前記領域識別手段で切り出された文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段と，を備えたことを特徴とする文書画像からのタイトル領域抽出装置。

【請求項２】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるか否かを判断する第１の工程と，前記第１の工程で一定のしきい値以上である場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第２の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項３】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字認識を実行し，該文字認識時に文字列矩形内の文字数を求める第１の工程と，文書のタイトルの文字数を用い，前記文字数と比較し，文字矩形数が所定値内であるか否かを判断する第２の工程と，前記第２の工程で，文字矩形数が所定値内である場合に，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項４】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，体言止めになっている領域であるかを判断する第２の工程と，前記第２の工程で体言止めになっている領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項５】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，タイトルに頻出する語尾の統計情報辞書と前記文字列領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断する第２の工程と，前記第２の工程の領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項６】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で特定のフォントを用いている文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項７】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項８】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形内の各文字矩形のアスペクト比を求める第１の工程と，前記アスペクト比に基づいて倍角文字であるかを判断する第２の工程と，前記倍角文字であると判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項９】　入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形に対して文字認識処理を実行する第１の工程と，前記文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める第２の工程と，前記合計値が前記文字矩形領域のほぼ半分であるかを判断する第３の工程と，前記第３の工程でほぼ半分であると判定された文字列領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第４の工程と，を含むことを特徴とする文書画像からのタイトル領域抽出方法。

【請求項１０】　前記タイトルらしさのポイント加算の可否判断に用いる基準値は，ユーザ単位の入力文書形式に合わせて学習して得られる最適値とし，可変・設定されることを特徴とする請求項２ないし９の何れか一つに記載の文書画像からのタイトル領域抽出方法。

【請求項１１】　文書画像を認識し，その結果に対して言語処理を行ってキーワードを抽出する第１の工程と，前記第１の工程で抽出されたキーワードと，請求項２ないし１０の何れか一つに記載の文書画像からのタイトル領域抽出方法に基づいて抽出したタイトルとを併記する第２の工程と，前記第２の工程で併記されたタイトルを用いて文書検索を実行する第３の工程と，を含むことを特徴とする文書検索方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は，ファクシミリやイメージスキャナ等の画像入力装置から入力された文書画像データのデータベースから，検索の利便性を向上させるために，文書内容を的確に表現するような文書中の領域をタイトル領域として抽出する文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法，並びに文書検索方法に関する。

【０００２】

【従来の技術】従来，文書画像を検索する際には，後の検索時の利便性を図るために，画像入力装置からの文書画像の入力とは別にオペレータが手作業で，その文書の内容を的確に表現するタイトル情報やキーワード情報を抽出／作成して付加したり，定形文書に対しては，文書中の特定の位置（文字列）をタイトル・キーワードとして切り出していた。

【０００３】また，非定形文書に対してレイアウト的特徴のみを用いてタイトルを抽出する参考技術文献が，例えば，特開平９－１３４４０６号公報の『文書画像からのタイトル抽出装置および方法』，特開平５－２７４４７１号公報の『イメージ文書のタイトル領域抽出処理方法』が開示されている。

【０００４】

【発明が解決しようとする課題】しかしながら，上記に示されるような従来の技術にあっては，オペレータによるタイトル情報やキーワード情報の付加は文書量が多くなるにしたがって作業量も増加するため，作業負担の増大化を招来させてしまう。また，特定の位置の自動切り出しを行うと，定形文書のみを対象とするので，非定形文書には利用することができず，利便性に欠けるといった問題点があった。

【０００５】従来より開示されている特開平９－１３４４０６号公報・特開平５－２７４４７１号公報にあっては，レイアウト的特徴にのみ注目してタイトル抽出を行っているため，文書内容を的確に表現するタイトルの的中率が必ずしも満足できるものではなく，後の文書検索等に支障をきたす等の問題点があった。

【０００６】本発明は，上記に鑑みてなされたものであって，特定の文書形式に依存せずにタイトル固有の特徴をポイントとして用いることにより，ポイント数の多い文字列領域をタイトルとして自動抽出し，タイトル抽出の的確性および文書検索時の利便性を向上させることを目的とする。

【０００７】

【課題を解決するための手段】上記の目的を達成するために，請求項１に係る文書画像からのタイトル領域抽出装置にあっては，画像入力装置から入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出装置において，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識手段と，前記領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と，前記文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と，前記領域識別手段で切り出された文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段と，を備えたものである。

【０００８】また，請求項２に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるか否かを判断する第１の工程と，前記第１の工程で一定のしきい値以上である場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第２の工程と，を含むものである。

【０００９】また，請求項３に係る文書画像からのタイ

トル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字認識を実行し，該文字認識時に文字列矩形内の文字数を求める第１の工程と，文書のタイトルの文字数を用い，前記文字数と比較し，文字矩形数が所定値内であるか否かを判断する第２の工程と，前記第２の工程で，文字矩形数が所定値内である場合に，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１０】また，請求項４に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，体言止めになっている領域であるかを判断する第２の工程と，前記第２の工程で体言止めになっている領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１１】また，請求項５に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域内の文字コードの認識結果に対して自然言語処理を実行する第１の工程と，前記第１の工程の結果，タイトルに頻出する語尾の統計情報辞書と前記文字列領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断する第２の工程と，前記第２の工程の領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１２】また，請求項６に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で特定のフォントを用いている文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１３】また，請求項７に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列領域に対してフォント識別処理を実行する第１の工程と，前記フォント識別処理の結果に基づいて，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する第２の工程と，前記第２の工程で判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１４】また，請求項８に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形内の各文字矩形のアスペクト比を求める第１の工程と，前記アスペクト比に基づいて倍角文字であるかを判断する第２の工程と，前記倍角文字であると判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第３の工程と，を含むものである。

【００１５】また，請求項９に係る文書画像からのタイトル領域抽出方法にあっては，入力された文書画像から文字列領域を矩形で切り出し，前記文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する文書画像からのタイトル領域抽出方法において，前記文字列矩形に対して文字認識処理を実行する第１の工程と，前記文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める第２の工程と，前記合計値が前記文字矩形領域のほぼ半分であるかを判断する第３の工程と，前記第３の工程でほぼ半分であると判定された文字列領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する第４の工程と，を含むものである。

【００１６】また，請求項１０に係る文書画像からのタイトル領域抽出方法にあっては，前記タイトルらしさのポイント加算の可否判断に用いる基準値は，ユーザ単位の入力文書形式に合わせて学習して得られる最適値とし，可変・設定されるものである。

【００１７】また，請求項１１に係る文書検索方法にあっては，文書画像を認識し，その結果に対して言語処理を行ってキーワードを抽出する第１の工程と，前記第１の工程で抽出されたキーワードと，請求項２ないし１０の何れか一つに記載の文書画像からのタイトル領域抽出

方法に基づいて抽出したタイトルとを併記する第２の工程と，前記第２の工程で併記されたタイトルを用いて文書検索を実行する第３の工程と，を含むものである。

【００１８】

【発明の実施の形態】以下，本発明の文書画像からのタイトル領域抽出装置およびタイトル領域抽出方法，並びに文書検索方法について添付図面を参照して説明する。

【００１９】図１は，本発明の実施の形態に係るタイトル領域抽出処理を行うシステム構成を示すブロック図である。図において，１０１はファクシミリやイメージスキャナ等の画像入力装置（図示せず）から入力された文書画像から文字列領域を矩形で切り出す領域識別手段としての領域識別部，１０２は領域識別部１０１の識別結果に基づいて文字認識を行う文字認識手段としての文字認識部，１０３は領域識別部１０１の識別結果に基づいてフォント識別を行うフォント識別手段としてのフォント識別部，１０４は文字認識部１０２の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段としての自然言語解析部，１０５は従来から用いられているセンタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段としてのポイント付部である。

【００２０】図１の構成において，画像入力装置（図示せず）から文書画像が入力されると，スキュー補正等の前処理を行い，領域識別部１０１により領域識別処理を実行し，文字列矩形の座標値・大きさの情報を得る。次いで，領域識別部１０１による領域識別処理の結果を用い，文字認識部１０２による文字認識，およびフォント識別部１０３によるフォント識別を行う。

【００２１】文字認識部１０２では各文字毎の文字コード・確信度，文字矩形の座標値・大きさがタイトルらしさのポイント付けとして得られる。また，フォント識別部１０３では各文字毎のフォント種別がタイトルらしさのポイント付けとして得られる。

【００２２】また，文字認識部１０２により得られる文字コードは，自然言語解析部１０４自然言語解析ルーチンにも供給され，自然言語的タイトルらしさ，つまり，体言止めになっている領域のタイトルらしさのポイントを与える。さらに，自然言語処理において，タイトルに頻出する語尾の統計情報辞書と文字領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域にタイトルらしさのポイントを与える。

【００２３】また，上述の各ポイントらしさのポイントに加え，従来から用いられているセンタリング処理・下線処理・文字列矩形の大きさ等も用いてタイトルらしさの合計ポイントを計算し，タイトルを識別する。

【００２４】次に，図３～図８に示すフローチャートを参照し，本発明の一連のタイトル抽出処理方法について順に説明する。なお，このタイトル抽出処理方は，図１の構成によって複数の組み合わせあるいは単独，あるいは選択的に行ってことができる。

【００２５】図３は，実施の形態に係る第１のタイトル抽出方法を示すフローチャートであり，文字コード識別の確信度が一定のしきい値以上であった場合にタイトルらしさのポイントを加算する例について示している。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ３０１），領域識別部１０１により文字列領域を識別する（Ｓ３０２）。続いて，上記文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるか否かを判断する（Ｓ３０３）。ここで，一定のしきい値以上であると判断した場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ３０４）。

【００２６】図４は，実施の形態に係る第２のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ４０１），領域識別部１０１により文字列領域を識別する（Ｓ４０２）。続いて，文字認識時に文字列矩形内の文字数を求める（Ｓ４０３）。そして，文書のタイトルの文字数を用い，上記文字数と比較し（Ｓ４０４），文字矩形数が所定値内であるか否かを判断する（Ｓ４０５）。ここで，文字矩形数が所定値内であると判断すると，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ４０６）。

【００２７】すなわち，文字列領域内の文字コード認識時に文字列矩形内の文字数を求め，別途辞書情報として文書のタイトルの文字数の統計を用いて比較し，タイトルらしい文字数の文字列矩形に対してタイトルらしさのポイントを与える。

【００２８】図５は，実施の形態に係る第３のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ５０１），領域識別部１０１により文字列領域を識別する（Ｓ５０２）。続いて，文字列領域内の文字コードの認識結果に対して自然言語処理を行い（Ｓ５０３），所定事項の領域，例えば，体言止めになっている領域か否かを判断する（Ｓ５０４）。ここで，所定事項の領域であると判断すると，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ５０５）。

【００２９】また，上述の言語処理において，タイトルに頻出する語尾の統計情報辞書と文字領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域にタイトルらしさのポイントを与えてもよい。

【００３０】図６は，実施の形態に係る第４のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ６０１），領域識別部１０１により文字列領域を識別する（Ｓ６０２）。続いて，フォント識別処理を行い（Ｓ６０３），所定のフォント（フォントスタイル）を含む領域である

か否かを判断する（Ｓ６０４）。つまり，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるか，あるいは，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断する。そして，これらの領域であると判断した場合に，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ６０５）。

【００３１】図７は，実施の形態に係る第５のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ７０１），領域識別部１０１により文字列領域を識別する（Ｓ７０２）。続いて，文字列矩形内の各文字矩形のアスペクト比を求め（Ｓ７０３），アスペクト比が横：縦＝２：１に近い値となる文字矩形が文字列矩形内の文字矩形数中の一定の割合以上を占めているか否かを判断する（Ｓ７０４）。ここで，一定以上の割合を占めていれば，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ７０５）。

【００３２】図８は，実施の形態に係る第６のタイトル抽出方法を示すフローチャートである。まず，文書入力装置（図示せず）から文書画像を入力し（Ｓ８０１），領域識別部１０１により文字列領域を識別する（Ｓ８０２）。続いて，文字認識処理を行い（Ｓ８０３），文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求める（Ｓ８０４）。そして，その合計値が文字矩形領域のほぼ半分であるか否かを判断する（Ｓ８０５）。ここで，合計値が文字矩形領域のほぼ半分であれば，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出する（Ｓ８０６）。

【００３３】ところで，上述した実施の形態において必要となるしきい値を固定せずに，各ユーザの入力する文書の対応させて学習し，各ユーザの使用する文書形式に対して最適なしきい値を可変的に求め，初期値から変更・設定してもよい。

【００３４】さらに，上述の如く求められる一時的なポイントを基に，図２に示すように二次的な組み合わせにより，倍角文字や均等割付けの判定を行い，それらに対してタイトルらしさのポイントを与えることも可能である。

【００３５】これを付言すると，文字コードの認識時に得られる文字列矩形内の各文字矩形領域の大きさを用い，文字矩形領域のアスペクト比を算出することによって倍角文字を判定し，該倍角文字を用いている文字列領域に対してタイトルらしさのポイントを与える。

【００３６】また，文字矩形領域とそれが属する文字列領域の大きさと，文字コードの認識時に得られる文字列矩形内の文字数とを用いて矩形内の文字密度を算出し，その値によって均等割付け判定を行う。そして，均等割付けされたと判定された文字列領域に対してタイトルらしさのポイントを与える。

【００３７】ところで，上述したタイトル領域抽出方法を用いて情報検索を行うことも実現可能である。図９は，実施の形態に係る情報検索方法を示すフローチャートである。まず，文書画像を認識し（Ｓ９０１），その結果に対して言語処理を行ってキーワードを抽出する（ＳＳ９０２）。さらに，上記抽出されたキーワードと，前述のタイトル領域抽出方法によって抽出したタイトルとを併記し（Ｓ９０３），その併記タイトルを用いて文書検索を実行する（Ｓ９０４）。これにより，検索時における利便性が向上する。

【００３８】

【発明の効果】以上説明したように，本発明に係る文書画像からのタイトル領域抽出装置（請求項１）によれば，入力された文書画像から文字列領域を矩形で切り出す領域識別手段を有し，その文字列領域の属性に基づいてタイトルらしさのポイント計算を実行し，タイトルを抽出する際に，領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の文字認識を行う文字認識手段と，領域識別手段で切り出された文字列矩形に対し，該文字列矩形内の各文字毎のフォント識別を行うフォント識別手段と，文字認識手段の認識結果で得られる文字コードに基づいて自然言語的タイトルらしさを解析する自然言語解析手段と，領域識別手段で切り出された文字列矩形に対し，センタリング・下線・文字矩形の大きさ等を用いてタイトルらしさのポイント付けを行うポイント付手段とを設け，特定の文書形式に依存せずにタイトル固有の特徴をポイント付けとして用いるため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させる装置を提供することができる。

【００３９】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項２）によれば，文字列領域内の文字コードを認識し，文字コード識別の確信度が一定のしきい値以上であるかを判断し，一定のしきい値以上である場合，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４０】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項３）によれば，文字列領域内の文字認識を実行し，該文字認識時に文字列矩形内の文字数を求め，文書のタイトルの文字数を用い，上記文字数と比較し，文字矩形数が所定値内であるかを判断し，文字矩形数が所定値内である場合，該当する文字列領域にタイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い

文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４１】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項４）によれば，文字列領域内の文字コードの認識結果に対して自然言語処理を実行し，そのの結果，体言止めになっている領域であるかを判断し，体言止めになっている領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４２】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項５）によれば，文字列領域内の文字コードの認識結果に対して自然言語処理を実行し，その結果，タイトルに頻出する語尾の統計情報辞書と文字列領域内の文字コード列とを比較し，高頻出度の語尾と一致するものを語尾に含む文字列領域であるかを判断し，その領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４３】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項６）によれば，文字列領域に対してフォント識別処理を実行し，その結果に基づいて，文字のフォントスタイルを判別し，特定のフォントを用いている文字領域であるかを判断し，特定のフォントを用いている文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４４】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項７）によれば，文字列領域に対してフォント識別処理を実行し，その結果に基づいて，フォントスタイル判別時に文書全体のフォントスタイルのヒストグラムを作成しておき，出現頻度の少ないフォントスタイルを用いている文字領域であるかを判断し，該判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４５】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項８）によれば，文字列矩形内の各文字矩形のアスペクト比を求め，そのアスペクト比に基づいて倍角文字であるかを判断し，倍角文字であると判断した文字領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４６】また，本発明に係る文書画像からのタイトル領域抽出方法（請求項９）によれば，文字列矩形に対して文字認識処理を実行し，文字認識処理によって空白文字以外認識された各文字矩形の横幅（縦書き時は縦幅）の合計値を求め，その合計値が文字矩形領域のほぼ半分であるかを判断し，ほぼ半分であると判定された文字列領域に対し，タイトルらしさのポイントを加算し，その合計値によりタイトル領域を決定・抽出するため，ポイント数の多い文字列領域をタイトルとして自動抽出することが実現し，かつタイトル抽出の的確性および文書検索時の利便性を向上させることができる。

【００４７】また，本発明に係る書画像からのタイトル領域抽出方法（請求項１０）によれば，請求項２ないし９の何れか一つに記載の文書画像からのタイトル領域抽出方法において，タイトルらしさのポイント加算の可否判断に用いる基準値を，ユーザ単位の入力文書形式に合わせて学習して得られる最適値を用いて可変・設定することにより，より的確なタイトルの自動抽出が実現する。

【００４８】また，本発明に係る文書検索方法（請求項１１）によれば，文書画像を文字認識し，その結果に対して言語処理を行って抽出されたキーワードと，請求項２ないし９の何れか一つに記載の文書画像からのタイトル領域抽出方法に基づいて抽出したタイトルとを併記し，該併記されたタイトル，すなわち，より的確なタイトルを用いて文書検索を実行するため，文書検索時における利便性が向上する。

【図面の簡単な説明】

【図１】本発明の実施の形態に係るタイトル領域抽出処理を行うシステム構成を示すブロック図である。

【図２】本発明の実施の形態に係るタイトル領域抽出処理に用いられるタイトルらしさのポイントうち，二次的に求められるものを示すブロック図である。

【図３】本発明の実施の形態に係る第１のタイトル抽出方法を示すフローチャートである。

【図４】本発明の実施の形態に係る第２のタイトル抽出方法を示すフローチャートである。

【図５】本発明の実施の形態に係る第３のタイトル抽出方法を示すフローチャートである。

【図６】本発明の実施の形態に係る第４のタイトル抽出方法を示すフローチャートである。

【図７】本発明の実施の形態に係る第５のタイトル抽出方法を示すフローチャートである。

【図８】本発明の実施の形態に係る第６のタイトル抽出

方法を示すフローチャートである。

【図９】本発明の実施の形態に係る情報検索方法を示すフローチャートである。

【符号の説明】

１０１　領域識別部

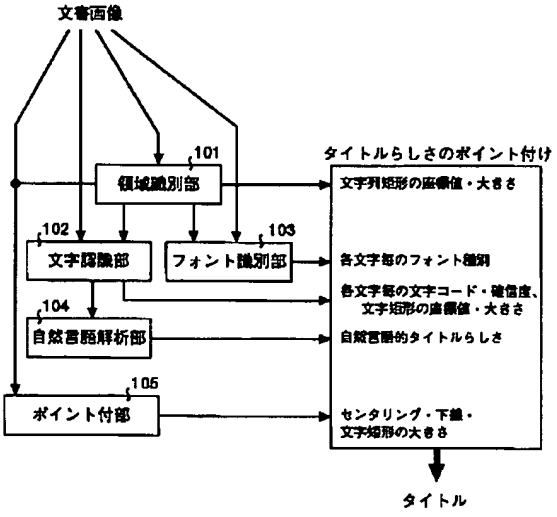
１０２　文字認識部

１０３　フォント識別部

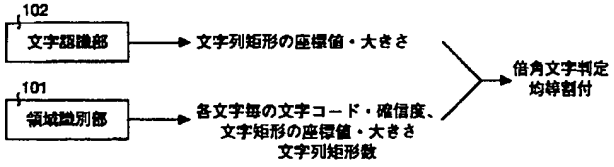
１０４　自然言語解析部

１０５　ポイント付部

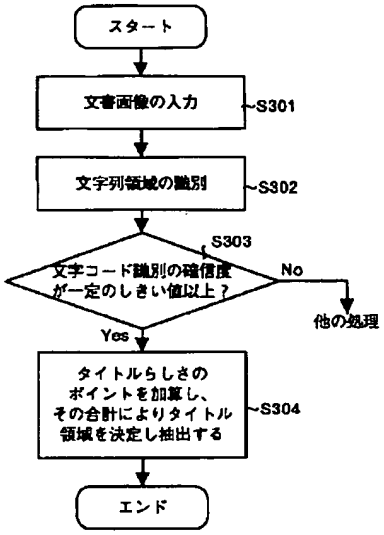
【図１】



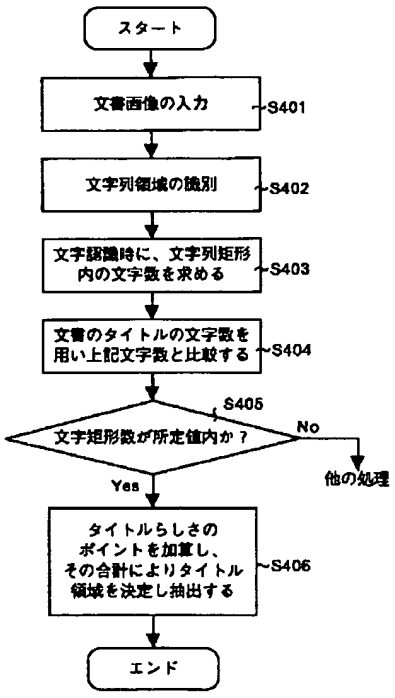
【図２】



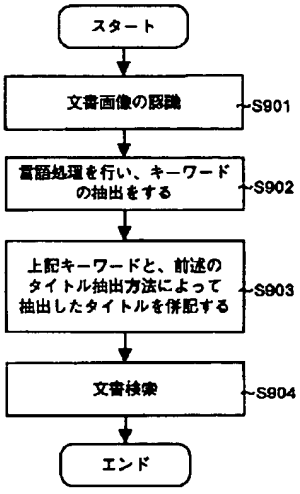
【図３】



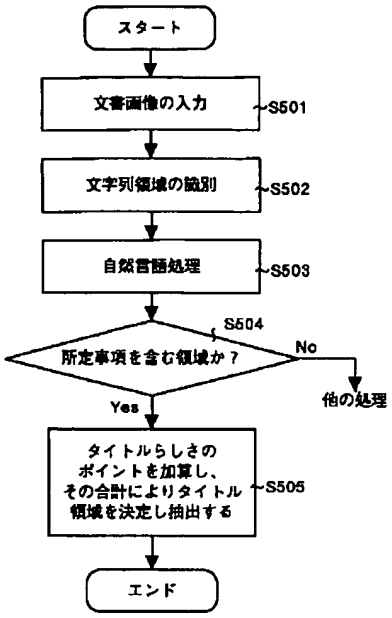
【図４】



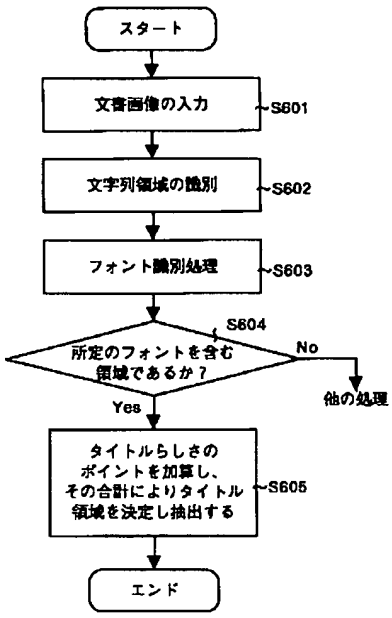
【図９】



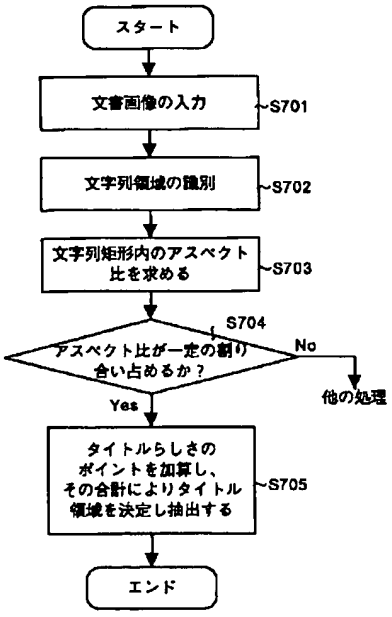
【図５】


スタート
文書画像の入力
S501
文字列領域の識別
S502
自然言語処理
S503
S504
所定事項を含む領域か？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S505
エンド

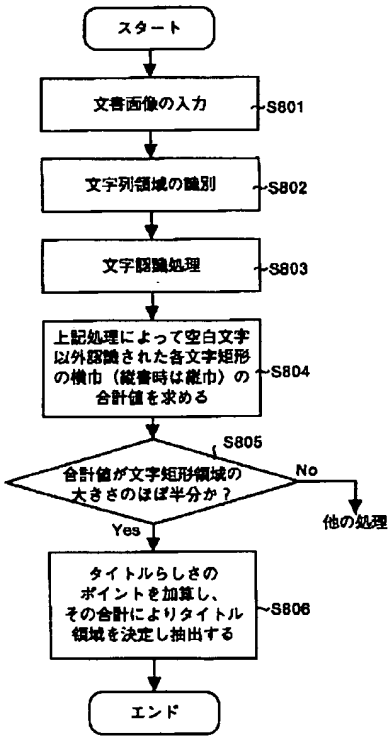
【図６】


スタート
文書画像の入力
S601
文字列領域の識別
S602
フォント識別処理
S603
S604
所定のフォントを含む
領域であるか？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S605
エンド

【図７】


スタート
文書画像の入力
S701
文字列領域の識別
S702
文字列矩形内のアスペクト
比を求める
S703
S704
アスペクト比が一定の割り
合い占めるか？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S705
エンド

【図８】


スタート
文書画像の入力
S801
文字列領域の識別
S802
文字認識処理
S803
上記処理によって空白文字
以外認識された各文字矩形
の横巾（縦書時は縦巾）の
合計値を求める
S804
S805
合計値が文字矩形領域の
大きさのほぼ半分か？
No
他の処理
Yes
タイトルらしさの
ポイントを加算し、
その合計によりタイトル
領域を決定し抽出する
S806
エンド

PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2000-148788**

(43)Date of publication of application : **30.05.2000**

(51)Int.Cl. **G06F 17/30**

G06F 17/27

G06K 9/20

(21)Application number : **10-328806**

(22)Date of filing : **05.11.1998**

(71)Applicant : **RICOH CO LTD**

(72)Inventor : **OUCHI SHIGEKI**

(54) DEVICE AND METHOD FOR EXTRACTING TITLE AREA FROM DOCUMENT IMAGE AND DOCUMENT RETRIEVING METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To improve the accuracy of title extraction and the convenience of document retrieval by analyzing the likelihood of a tile which is linguistically natural according to a character code obtained as a recognition result and giving points to the title likelihood by using centering, underscoring, the size of a character rectangle, etc.

SOLUTION: A character recognition part 102 obtains the character code, accuracy, coordinate values of the character rectangle, and size of each character as points as title likelihood. A font identification part 103 obtains the font kind of each character as points as the title likelihood. The character code is supplied to a natural language analyzing routine of a natural language analysis part 104 to give points as natural language title likelihood and the points of title likelihood are given to a character string area having a word ending matching the word ending appearing frequency in a title through a natural language process. Further, total points of title likelihood are calculated by using a centering process, an underscoring process, the size of the character string rectangle, etc., to identify the title.

LEGAL STATUS

[Date of request for examination] 08.11.2002

[Date of sending the examiner's decision of rejection] 07.12.2004

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

CLAIMS

[Claim(s)]

[Claim 1] A string area from the document image inputted from the picture input device In the title field extractor from the document image which has the field discernment means started with a rectangle, performs point count of title-likeness based on the attribute of said string area, and extracts a title A character recognition means to perform character recognition in this character string rectangle to the character string rectangle started with said field discernment means, A font discernment means to perform font discernment for every alphabetic character in this character string rectangle to the character string rectangle started with said field discernment means, A natural language analysis means to analyze natural language-title-likeness based on the character code which it may be as a result of [of said character recognition means] recognition, The title field extractor from the document image characterized by having a means with the point to perform point attachment of title-likeness using the magnitude of centering, an underline, and an alphabetic character rectangle etc., to the character string rectangle started with said field discernment means.

[Claim 2] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which recognizes the character code in said string area, and judges whether it is more than a threshold with the fixed reliability of character code discernment, The title field extract approach from the document image characterized by including

the 2nd process which adds the point of title-likeness, and determines and extracts a title field with the total value at said 1st process when it is more than a fixed threshold.

[Claim 3] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs character recognition in said string area, and asks for the number of alphabetic characters in a character string rectangle at the time of this character recognition, As compared with said number of alphabetic characters, using the number of alphabetic characters of the title of a document at the 2nd process which judges whether the number of alphabetic character rectangles is in a predetermined value, and said 2nd process The title field extract approach from the document image characterized by including the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value when the number of alphabetic character rectangles is in a predetermined value.

[Claim 4] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs natural language processing to the recognition result of the character code in said string area, The 2nd process which judges whether it is the field which is a substantives stop as a result of said 1st process, The title field extract approach from the document image characterized by including the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the field which is a substantives stop at said 2nd process.

[Claim 5] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs natural language processing to the recognition result of the character code in said string area, The 2nd process which judges whether it is the string area which compares with the character code train in said string area the statistical information dictionary of the ending which occurs frequently in a title as a result of said 1st process, and contains the ending and the match of high frequency in the ending, The title field extract approach from the document image characterized by including the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the field of said 2nd process.

[Claim 6] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs font discernment processing to said string area, The 2nd process which judges whether it is the alphabetic character field which distinguishes the font style of an alphabetic character and uses the specific font based on the

result of said font discernment processing, The title field extract approach from the document image characterized by including the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the alphabetic character field which uses the specific font at said 2nd process.

[Claim 7] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs font discernment processing to said string area, The 2nd process which judges whether it is the alphabetic character field which creates the histogram of the font style of the whole document and uses the font style with little frequency of occurrence based on the result of said font discernment processing at the time of font style distinction, The title field extract approach from the document image characterized by including the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the alphabetic character field judged at said 2nd process.

[Claim 8] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which asks for the aspect ratio of each alphabetic character rectangle in said character string rectangle, The 2nd process which judges whether it is a double width character based on said aspect ratio, The title field extract approach from the document image characterized by including the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the alphabetic character field judged to be said double width character.

[Claim 9] In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs character recognition processing to said character string rectangle, and the 2nd process which calculates the total value of the breadth (it is a dip at the time of columnar writing) of each alphabetic character rectangle recognized by said character recognition processing except the null character, The 3rd process said total value judges [of said alphabetic character rectangle field] it to be whether it is one half mostly, The title field extract approach from the document image characterized by including the 4th process which adds the point of title-likeness, and determines and extracts a title field with the total value to the string area judged that is one half mostly at said 3rd process.

[Claim 10] The reference value used for propriety decision of point addition of said title-likeness is the title field extract approach from claim 2 which considers as the optimum value learned and acquired according to the input-statement document format of a user unit, and is characterized by adjustable and set up thru/or the document image of any of 9, or one publication.

[Claim 11] The 1st process which recognizes a document image, performs language processing to the result, and extracts a keyword, The 2nd process which writes together the keyword extracted at said 1st process, and claim 2 thru/or any of 10 or the title extracted [one] based on the title field extract approach from the document image of a publication, The document-retrieval approach characterized by including the 3rd process which performs a document retrieval using the title written together at said 2nd process.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention] This invention relates to the document-retrieval approach at the title field extractor from a document image and the title field extract approach of extracting the field in a document which expresses the contents of a document exactly as a title field from the database of document image data inputted from picture input devices, such as facsimile and an image scanner, in order to raise the convenience of retrieval, and a list.

[0002]

[Description of the Prior Art] In case a document image be search conventionally , in order to plan convenience at the time of next retrieval , extracted / created the title information and keyword information which an operator be handicraft apart from the input of the document image from a picture input device , and express the contents of the document exactly , and they be added , and the specific location in a document (character string) be started as a title keyword to the fixed form document .

[0003] Moreover, "the title extractor from a document image and approach" of JP,9-134406,A, and the "title field extract art of an image document" of JP,5-274471,A are indicated for the reference technical reference which extracts a title only using the layout-description to a non-fixed form document.

[0004]

[Problem(s) to be Solved by the Invention] However, since rating also increases as the amount of documents increases, addition of the title information by the operator or keyword information will make increase-ization of an activity burden invite, if it is in a Prior art as shown above. Moreover, since it was aimed only at the fixed form document when automatic logging of a specific location was performed, it could not use for a non-fixed form document, but there was a trouble that convenience was missing.

[0005] If it was in JP,9-134406,A and JP,5-274471,A currently indicated

conventionally, since the title extract was performed only paying attention to the layout-description, the hitting ratio of the title which expresses the contents of a document exactly cannot necessarily be satisfied, and there were troubles, such as causing trouble to a next document retrieval etc.

[0006] By being made in view of the above and using the description of a title proper as the point, without being dependent on a specific document format, this invention makes a title a string area with many point sizes, carries out automatic extracting, and aims at raising the exact nature of a title extract, and the convenience at the time of a document retrieval.

[0007]

[Means for Solving the Problem] If it is in a title field extractor from the document image concerning claim 1 in order to attain the above-mentioned purpose A string area from the document image inputted from the picture input device In the title field extractor from the document image which has the field discernment means started with a rectangle, performs point count of title-likeness based on the attribute of said string area, and extracts a title A character recognition means to perform character recognition in this character string rectangle to the character string rectangle started with said field discernment means, A font discernment means to perform font discernment for every alphabetic character in this character string rectangle to the character string rectangle started with said field discernment means, A natural language analysis means to analyze natural language-title-likeness based on the character code which it may be as a result of [of said character recognition means] recognition, It has a means with the point to perform point attachment of title-likeness using the magnitude of centering, an underline, and an alphabetic character rectangle etc., to the character string rectangle started with said field discernment means.

[0008] Moreover, if it is in the title field extract approach from the document image concerning claim 2 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which recognizes the character code in said string area, and judges whether it is more than a threshold with the fixed reliability of character code discernment, When it is more than a fixed threshold at said 1st process, the point of title-likeness is added and the 2nd process which determines and extracts a title field with the total value is included.

[0009] Moreover, if it is in the title field extract approach from the document image concerning claim 3 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs character recognition in said string area, and asks for the number of alphabetic characters in a character string rectangle at the

time of this character recognition, As compared with said number of alphabetic characters, using the number of alphabetic characters of the title of a document at the 2nd process which judges whether the number of alphabetic character rectangles is in a predetermined value, and said 2nd process When the number of alphabetic character rectangles is in a predetermined value, the point of title-likeness is added and the 3rd process which determines and extracts a title field with the total value is included.

[0010] Moreover, if it is in the title field extract approach from the document image concerning claim 4 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs natural language processing to the recognition result of the character code in said string area, The 2nd process which judges whether it is the field which is a substantives stop, and the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the field which is a substantives stop at said 2nd process are included as a result of said 1st process.

[0011] Moreover, if it is in the title field extract approach from the document image concerning claim 5 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs natural language processing to the recognition result of the character code in said string area, The 2nd process which judges whether it is the string area which compares with the character code train in said string area the statistical information dictionary of the ending which occurs frequently in a title as a result of said 1st process, and contains the ending and the match of high frequency in the ending, To the field of said 2nd process, the point of title-likeness is added and the 3rd process which determines and extracts a title field with the total value is included.

[0012] Moreover, if it is in the title field extract approach from the document image concerning claim 6 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs font discernment processing to said string area, The 2nd process which judges whether it is the alphabetic character field which distinguishes the font style of an alphabetic character and uses the specific font based on the result of said font discernment processing, To the alphabetic character field which uses the specific font at said 2nd process, the point of title-likeness is added and the 3rd process which determines and extracts a title field with the total value is included.

[0013] Moreover, if it is in the title field extract approach from the document image

concerning claim 7 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs font discernment processing to said string area, The 2nd process which judges whether it is the alphabetic character field which creates the histogram of the font style of the whole document and uses the font style with little frequency of occurrence based on the result of said font discernment processing at the time of font style distinction, To the alphabetic character field judged at said 2nd process, the point of title-likeness is added and the 3rd process which determines and extracts a title field with the total value is included.

[0014] Moreover, if it is in the title field extract approach from the document image concerning claim 8 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which asks for the aspect ratio of each alphabetic character rectangle in said character string rectangle, The 2nd process which judges whether it is a double width character based on said aspect ratio, and the 3rd process which adds the point of title-likeness, and determines and extracts a title field with the total value to the alphabetic character field judged to be said double width character are included.

[0015] Moreover, if it is in the title field extract approach from the document image concerning claim 9 In the title field extract approach from a document image of starting a string area with a rectangle from the inputted document image, performing point count of title-likeness based on the attribute of said string area, and extracting a title The 1st process which performs character recognition processing to said character string rectangle, and the 2nd process which calculates the total value of the breadth (it is a dip at the time of columnar writing) of each alphabetic character rectangle recognized by said character recognition processing except the null character, Said total value includes the 3rd process which judges [of said alphabetic character rectangle field] whether it is one half mostly, and the 4th process which adds the point of title-likeness, and determines and extracts a title field with the total value to the string area judged that is one half mostly at said 3rd process.

[0016] Moreover, if it is in the title field extract approach from the document image concerning claim 10, the reference value used for propriety decision of point addition of said title-likeness is made into the optimum value learned and acquired according to the input-statement document format of a user unit, and are adjustable and a thing set up.

[0017] Moreover, if it is in the document-retrieval approach concerning claim 11 The 1st process which recognizes a document image, performs language processing to the result, and extracts a keyword, The 2nd process which writes together the keyword extracted at said 1st process, and claim 2 thru/or any of 10 or the title extracted

[one] based on the title field extract approach from the document image of a publication, The 3rd process which performs a document retrieval using the title written together at said 2nd process is included.
[0018]
[Embodiment of the Invention] Hereafter, the document-retrieval approach is explained to the title field extractor from the document image of this invention and the title field extract approach, and a list with reference to an accompanying drawing.
[0019] Drawing 1 is the block diagram showing the system configuration which performs title field extract processing concerning the gestalt of operation of this invention. In drawing, 101 as a field discernment means which starts a string area with a rectangle from the document image inputted from picture input devices (not shown), such as facsimile and an image scanner The ****** discernment section and 102 as a character recognition means to perform character recognition based on the discernment result of the field discernment section 101 ********** and 103 as a font discernment means to perform font discernment based on the discernment result of the field discernment section 101 The ** font discernment section and 104 as a natural language analysis means to analyze natural language-title-likeness based on the character code which it may be as a result of [of the character recognition section 102] recognition The ********** analysis section and 105 are the sections with the point as a means with the point which perform point attachment of title-likeness using the magnitude of the centering, underline, and alphabetic character rectangle used from the former etc.
[0020] In the configuration of drawing 1 , if a document image is inputted from a picture input device (not shown), skew correction etc. will be pretreated, field discernment processing will be performed by the field discernment section 101, and the information on the coordinate value and magnitude of a character string rectangle will be acquired. Subsequently, character recognition by the character recognition section 102 and font discernment by the font discernment section 103 are performed using the result of the field discernment processing by the field discernment section 101.
[0021] In the character recognition section 102, the coordinate value and magnitude of the character code and reliability for every alphabetic character, and an alphabetic character rectangle are obtained as point attachment of title-likeness. Moreover, in the font discernment section 103, the font classification for every alphabetic character is obtained as point attachment of title-likeness.
[0022] Moreover, the character code obtained by the character recognition section 102 is supplied also to a natural language analysis section 104 natural-language analyzer, and gives the point of natural language-title-likeness, i.e., title-likeness of the field which is a substantives stop. Furthermore, in natural language processing, the statistical information dictionary of the ending which occurs frequently in a title is compared with the character code train in an alphabetic character field, and the point

of title-likeness is given to the string area which contains the ending and the match of high frequency in the ending.

[0023] Moreover, the sum total point of title-likeness is calculated by using the magnitude of the centering processing, the underline processing, and the character string rectangle used from the former etc. in addition to the point of each above-mentioned point-likeness, and a title is identified.

[0024] Next, with reference to the flow chart shown in drawing 3 - drawing 8 , a series of title extract arts of this invention are explained in order. In addition, by the configuration of drawing 1 , the method of this title extract processing is performed on two or more combination, independent, or selection targets, and can do things.

[0025] Drawing 3 is a flow chart which shows the 1st title extract approach concerning the gestalt of operation, and when the reliability of character code discernment is more than a fixed threshold, it shows the example adding the point of title-likeness. First, a document image is inputted from a document input unit (not shown) (S301), and a string area is identified by the field discernment section 101 (S302). Then, the character code in the above-mentioned string area is recognized, and it judges whether it is more than a threshold with the fixed reliability of character code discernment (S303). Here, when it is judged that it is more than a fixed threshold, the point of title-likeness is added, and the total value determines and extracts a title field (S304).

[0026] Drawing 4 is a flow chart which shows the 2nd title extract approach concerning the gestalt of operation. First, a document image is inputted from a document input unit (not shown) (S401), and a string area is identified by the field discernment section 101 (S402). Then, it asks for the number of alphabetic characters in a character string rectangle at the time of character recognition (S403). And as compared with the above-mentioned number of alphabetic characters (S404), it judges whether the number of alphabetic character rectangles is in a predetermined value using the number of alphabetic characters of the title of a document (S405). Here, if it judges that the number of alphabetic character rectangles is in a predetermined value, the point of title-likeness will be added, and the total value will determine and extract a title field (S406).

[0027] That is, at the time of the character code recognition in a string area, it asks for the number of alphabetic characters in a character string rectangle, and compares separately, using statistics of the number of alphabetic characters of the title of a document as dictionary information, and the point of title-likeness is given to the character string rectangle of the number of alphabetic characters appropriate for a title.

[0028] Drawing 5 is a flow chart which shows the 3rd title extract approach concerning the gestalt of operation. First, a document image is inputted from a document input unit (not shown) (S501), and a string area is identified by the field discernment section 101 (S502). Then, natural language processing is performed to

the recognition result of the character code in a string area (S503), and it judges whether it is the field of a predetermined matter, for example, the field which is a substantives stop, (S504). Here, if it judges that it is the field of a predetermined matter, the point of title-likeness will be added, and the total value will determine and extract a title field (S505).

[0029] Moreover, in above-mentioned language processing, the statistical information dictionary of the ending which occurs frequently in a title is compared with the character code train in an alphabetic character field, and the point of title-likeness may be given to the string area which contains the ending and the match of high frequency in the ending.

[0030] Drawing 6 is a flow chart which shows the 4th title extract approach concerning the gestalt of operation. First, a document image is inputted from a document input unit (not shown) (S601), and a string area is identified by the field discernment section 101 (S602). Then, font discernment processing is performed (S603) and it judges whether it is a field containing a predetermined font (font style) (S604). That is, the font style of an alphabetic character is distinguished, the histogram of the font style of the whole document is created at the time of whether it is the alphabetic character field which uses the specific font, and font style distinction, and it judges whether it is the alphabetic character field which uses the font style with little frequency of occurrence. And when it is judged that they are these fields, the point of title-likeness is added, and the total value determines and extracts a title field (S605).

[0031] Drawing 7 is a flow chart which shows the 5th title extract approach concerning the gestalt of operation. First, a document image is inputted from a document input unit (not shown) (S701), and a string area is identified by the field discernment section 101 (S702). Then, it asks for the aspect ratio of each alphabetic character rectangle in a character string rectangle (S703), and the alphabetic character rectangle from which an aspect ratio serves as a value near horizontal:length =2:1 judges whether it has accounted more than for the fixed rate in the number of alphabetic character rectangles in a character string rectangle (S704). Here, if it has accounted for the rate more than fixed, the point of title-likeness will be added, and the total value will determine and extract a title field (S705).

[0032] Drawing 8 is a flow chart which shows the 6th title extract approach concerning the gestalt of operation. First, a document image is inputted from a document input unit (not shown) (S801), and a string area is identified by the field discernment section 101 (S802). Then, character recognition processing is performed (S803) and the total value of the breadth (it is a dip at the time of columnar writing) of each alphabetic character rectangle recognized by character recognition processing except the null character is calculated (S804). And the total value judges [of an alphabetic character rectangle field] whether it is one half mostly (S805). Here, if an alphabetic character rectangle field is one half mostly, total value will add

the point of title-likeness, and will determine and extract a title field with the total value (S806).

[0033] By the way, the document which each user inputs may make it correspond without fixing the threshold which is needed in the gestalt of operation mentioned above, it may learn, the optimal threshold may be calculated in adjustable from the document format which each user uses, and you may change and set up from initial value.

[0034] Furthermore, it is also possible like **** to perform the judgment of a double width character or an equal space, and to give the point of title-likeness to them with a secondary combination, based on the temporary point called for, as shown in drawing 2 .

[0035] If this is added, using each alphabetic character rectangle area size in the character string rectangle obtained at the time of recognition of a character code, by computing the aspect ratio of an alphabetic character rectangle field, a double width character will be judged and the point of title-likeness will be given to the string area which uses this double width character.

[0036] Moreover, the character density in a rectangle is computed using the magnitude of an alphabetic character rectangle field and the string area where it belongs, and the number of alphabetic characters in the character string rectangle obtained at the time of recognition of a character code, and the value performs an equal space judging. And the point of title-likeness is given to the string area judged that the equal space was carried out.

[0037] By the way, it is also realizable to perform information retrieval using the title field extract approach mentioned above. Drawing 9 is a flow chart which shows the information retrieval approach concerning the gestalt of operation. First, a document image is recognized (S901), language processing is performed to the result, and a keyword is extracted (SS902). Furthermore, the keyword by which the extract was carried out [above-mentioned], and the title extracted by the above-mentioned title field extract approach are written together (S903), and a document retrieval is performed using the writing-together title (S904). Thereby, the convenience at the time of retrieval improves.

[0038]

[Effect of the Invention] As explained above, according to the title field extractor (claim 1) from the document image concerning this invention A string area from the inputted document image The field discernment means started with a rectangle A character recognition means to perform character recognition in this character string rectangle to the character string rectangle started with the field discernment means in case it has, point count of title-likeness is performed based on the attribute of the string area and a title is extracted, A font discernment means to perform font discernment for every alphabetic character in this character string rectangle to the character string rectangle started with the field discernment means, A natural

language analysis means to analyze natural language-title-likeness based on the character code which it may be as a result of [of a character recognition means] recognition, The description of a title proper is considered as point attachment, without establishing a means with the point to perform point attachment of title-likeness using the magnitude of centering, an underline, and an alphabetic character rectangle etc., to the character string rectangle started with the field discernment means, and being dependent on a specific document format. Since it uses, the equipment which it is realized that it carries out automatic extracting, using a string area with many point sizes as a title, and raises the exact nature of a title extract and the convenience at the time of a document retrieval can be offered.

[0039] Moreover, according to the title field extract approach (claim 2) from the document image concerning this invention The character code in a string area is recognized and it judges whether it is more than a threshold with the fixed reliability of character code discernment, and the point of title-likeness is added when it is more than a fixed threshold. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0040] Moreover, according to the title field extract approach (claim 3) from the document image concerning this invention The character recognition in a string area It performs and asks for the number of alphabetic characters in a character string rectangle at the time of this character recognition, as compared with the above-mentioned number of alphabetic characters, it judges whether the number of alphabetic character rectangles is in a predetermined value using the number of alphabetic characters of the title of a document, and when the number of alphabetic character rectangles is in a predetermined value, the point of title-likeness is added to the corresponding string area. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0041] Moreover, according to the title field extract approach (claim 4) from the document image concerning this invention Natural language processing is performed to the recognition result of the character code in a string area, it judges whether it is the field which is a substantives stop as a result of ****, and the point of title-likeness is added to the field which is a substantives stop. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0042] Moreover, according to the title field extract approach (claim 5) from the document image concerning this invention As opposed to the recognition result of the

character code in a string area Natural language processing The statistical information dictionary of the ending which performs, consequently occurs frequently in a title is compared with the character code train in a string area, it judges whether it is the string area which contains the ending and the match of high frequency in the ending, and the point of title-likeness is added to the field. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0043] Moreover, according to the title field extract approach (claim 6) from the document image concerning this invention Font discernment processing is performed to a string area, the font style of an alphabetic character is distinguished based on the result, it judges whether it is the alphabetic character field which uses the specific font, and the point of title-likeness is added to the alphabetic character field which uses the specific font. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0044] Moreover, according to the title field extract approach (claim 7) from the document image concerning this invention As opposed to a string area Font discernment processing As opposed to the alphabetic character field which performs, creates the histogram of the font style of the whole document based on the result at the time of font style distinction, judged whether it was the alphabetic character field which uses the font style with little frequency of occurrence, and was this judged the point of title-likeness Since it adds and the total value determines and extracts a title field, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0045] Moreover, according to the title field extract approach (claim 8) from the document image concerning this invention The point of title-likeness is added to the alphabetic character field which asked for the aspect ratio of each alphabetic character rectangle in a character string rectangle, judged whether it was a double width character based on the aspect ratio, and was judged to be a double width character. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0046] Moreover, according to the title field extract approach (claim 9) from the document image concerning this invention Character recognition processing is performed to a character string rectangle. By character recognition processing The total value of the breadth (it is a dip at the time of columnar writing) of each

alphabetic character rectangle recognized except the null character is calculated, the total value judges [of an alphabetic character rectangle field] whether it is one half mostly, and the point of title-likeness is added to the string area judged that is one half mostly. With the total value Since a title field is determined and extracted, it can be realized that it carries out automatic extracting, using a string area with many point sizes as a title, and the exact nature of a title extract and the convenience at the time of a document retrieval can be raised.

[0047] Moreover, according to the title field extract approach (claim 10) from the paintings-and-calligraphic-works image concerning this invention, in the title field extract approach from claim 2 thru/or the document image of any of 9, or one publication, automatic extracting of a more exact title is realized adjustable and by setting up using the optimum value which learns the reference value used for propriety decision of point addition of title-likeness according to the input-statement document format of a user unit, and is acquired.

[0048] Moreover, the keyword which was extracted by carrying out character recognition of the document image, and performing language processing to the result according to the document-retrieval approach (claim 11) concerning this invention, In order to write together claim 2 thru/or any of 9, or the title extracted [one] based on the title field extract approach from the document image of a publication and to perform a document retrieval using the this written-together title, i.e., a more exact title, the convenience at the time of a document retrieval improves.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] It is the block diagram showing the system configuration which performs title field extract processing concerning the gestalt of operation of this invention.

[Drawing 2] the point of title-likeness used for the title field extract processing concerning the gestalt of operation of this invention -- it is the block diagram showing what called for secondarily inside.

[Drawing 3] It is the flow chart which shows the 1st title extract approach concerning the gestalt of operation of this invention.

[Drawing 4] It is the flow chart which shows the 2nd title extract approach concerning the gestalt of operation of this invention.

[Drawing 5] It is the flow chart which shows the 3rd title extract approach concerning the gestalt of operation of this invention.

[Drawing 6] It is the flow chart which shows the 4th title extract approach concerning the gestalt of operation of this invention.

[Drawing 7] It is the flow chart which shows the 5th title extract approach concerning the gestalt of operation of this invention.
[Drawing 8] It is the flow chart which shows the 6th title extract approach concerning the gestalt of operation of this invention.
[Drawing 9] It is the flow chart which shows the information retrieval approach concerning the gestalt of operation of this invention.
[Description of Notations]
101 Field Discernment Section
102 Character Recognition Section
103 Font Discernment Section
104 Natural Language Analysis Section
105 Section with Point